Министерство образования Республики Беларусь

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра ИиТП

Тема реферата:

«МЕТОДЫ ОЦЕНИВАНИЯ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ, КРОССВАЛИДАЦИЯ»

Выполнил: Карпик Сергей Эдуардович
магистрант
заочного обучения
кафедры информатики и технологий
программирования
группа № 556241
контактный телефон: +375297546811
e-mail: serg.karpik@gmail.com

Проверил: Кандидат технических наук
Боброва Наталья Леонидовна

Минск 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 3
1. ОБЩАЯ ХАРАКТЕРИСТИКА МЕТОДА ОЦЕНИВАНИЯ ОБОЩАЮЩЕЙ СПОСОБНОСТИ 5
2. АЛГОРИТМ РЕАЛИЗАЦИИ МЕТОДА 8
3. КРОССВАЛИДАЦИЯ 10
4. СРАВНЕНИЯ С ДРУГИМИ МЕТОДАМИ АНАЛИЗА 12
5. СИЛЬНЫЕ И СЛАБЫЕ СТОРОНЫ МЕТОДА 15
6. АНАЛИЗ РЕЗУЛЬТАТОВ И ОБСУЖДЕНИЕ 16
ЗАКЛЮЧЕНИЕ 17

ВВЕДЕНИЕ

Методы оценивания обобщающей способности занимают важное место в современной аналитике данных и машинном обучении, поскольку позволяют определить, насколько построенная модель способна работать с новыми, ранее не встречавшимися данными. Главная цель построения любой модели — не просто запомнить обучающую выборку, а выявить закономерности, которые сохраняются и за её пределами. Именно поэтому проблема оценки обобщающей способности является одной из ключевых при разработке и внедрении интеллектуальных систем.

Одним из наиболее распространённых и эффективных способов такой оценки является кросс-валидация — метод, позволяющий объективно проверить качество модели, многократно разделяя исходные данные на обучающие и тестовые подмножества. Этот подход обеспечивает более точное понимание того, как модель поведёт себя при работе с реальными данными, и помогает избежать переобучения, которое нередко возникает при избыточной подгонке параметров под обучающую выборку.

Кросс-валидация и другие методы оценки обобщающей способности находят широкое применение в самых разных областях — от экономики и инженерии до медицины и биоинформатики. Они позволяют исследователям не только сравнивать различные алгоритмы, но и подбирать оптимальные параметры моделей, обеспечивая устойчивость и надёжность принимаемых решений.

В то же время использование этих методов требует внимательного подхода к выбору стратегии разделения данных, интерпретации результатов и оценке статистической значимости полученных показателей. Ошибки на этом этапе могут привести к неверным выводам о качестве модели и, как следствие, к неэффективным практическим решениям.

Таким образом, изучение методов оценивания обобщающей способности и принципов кросс-валидации является важным этапом подготовки специалистов, работающих в области анализа данных, статистики и машинного обучения. Понимание данных подходов способствует формированию системного взгляда на процесс построения моделей и повышает качество проводимых исследований.

Цель данного реферата – рассмотреть основные методы оценивания обобщающей способности моделей и подробно изучить принципы кросс-валидации как одного из наиболее универсальных инструментов проверки качества моделей.

Для достижения поставленной цели в работе решаются следующие задачи:

1. Изучить понятие обобщающей способности модели и её роль в машинном обучении,
2. Рассмотреть существующие методы оценки качества моделей,
3. Подробно описать принцип и разновидности кросс-валидации,
4. Сравнить эффективность кросс-валидации с другими методами оценки,
5. Проанализировать преимущества и ограничения кросс-валидации в практических применениях;

Результатом выполнения реферата станет формирование целостного представления о том, как методы оценивания обобщающей способности позволяют объективно судить о качестве модели и обеспечивают её надёжную работу в реальных условиях применения.

1. ОБЩАЯ ХАРАКТЕРИСТИКА МЕТОДА ОЦЕНИВАНИЯ ОБОЩАЮЩЕЙ СПОСОБНОСТИ

Обобщающая способность модели — это её свойство правильно обрабатывать новые, ранее не встречавшиеся данные, отражая реальные закономерности, а не случайные особенности обучающей выборки. Иными словами, модель с высокой обобщающей способностью демонстрирует устойчивые результаты как на обучающих данных, так и на данных, не участвовавших в обучении.

Необходимость оценки обобщающей способности возникает из-за проблемы переобучения (overfitting), когда модель чрезмерно подстраивается под обучающие данные, теряя способность делать корректные прогнозы на новых наблюдениях. Противоположная ситуация — недообучение (underfitting), при котором модель не способна уловить даже основные закономерности, содержащиеся в данных. Таким образом, правильное оценивание обобщающей способности позволяет найти баланс между сложностью модели и её точностью.

Методы оценивания обобщающей способности направлены на измерение того, насколько хорошо модель, обученная на одной части данных, способна предсказывать результаты на другой части, не использованной в обучении. Основная идея заключается в разделении исходного набора данных на подмножества:

– обучающее (training set) – используется для подбора параметров модели,

– тестовое (test set) – применяется для проверки качества предсказаний,

– иногда выделяется и валидационное (validation set) — для настройки гиперпараметров и выбора оптимальной архитектуры модели;

Наиболее простым способом оценки обобщающей способности является разделение данных на две части — обучающую и тестовую выборки (метод hold-out). Однако при небольших объёмах данных этот подход может давать нестабильные результаты, поскольку качество оценки сильно зависит от конкретного способа разбиения.

Для повышения надёжности оценки разработаны более сложные методы, такие как кросс-валидация (cross-validation), leave-one-out (исключение одного наблюдения) и бутстреп-методы (bootstrap). Они позволяют многократно использовать доступные данные, чередуя обучающие и тестовые подмножества, что обеспечивает более точное и устойчивое измерение качества модели.

Кроме того, оценка обобщающей способности связана с выбором метрики качества — числового показателя, отражающего точность или ошибку модели. Наиболее распространённые метрики включают среднеквадратичную ошибку (MSE), среднюю абсолютную ошибку (MAE), точность (accuracy), полноту (recall), F1-меру и другие, в зависимости от типа решаемой задачи.

Таким образом, методы оценивания обобщающей способности представляют собой фундаментальный инструмент анализа эффективности моделей машинного обучения. Их использование обеспечивает объективную проверку полученных результатов, способствует выявлению переобучения и помогает выбрать оптимальную модель, способную надёжно работать с реальными данными.

С математической точки зрения, процесс оценивания обобщающей способности можно представить как задачу минимизации обобщающей ошибки (generalization error), которая определяется как математическое ожидание ошибки модели на всей совокупности возможных данных [1] (1.1):

$$E_{gen} = \mathrm{E}_{(x,y)\sim Pdata}[L(f(x), y)] \quad (1.1)$$

где:

– E_{gen} – ожидаемая (обобщающая) ошибка модели,

– P_{data} – истинное распределение данных,

– $f(x)$ – предсказание модели,

– y – реальное значение,

– $L(f(x), y)$ – функция потерь (ошибки);

Так как истинное распределение Pdata обычно неизвестно, на практике обобщающую ошибку оценивают с помощью выборочной оценки ошибки (empirical error), получаемой на тестовой или валидационной выборке. Она вычисляется по формуле [1] (1.2):

$$\hat{E} = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) \quad (1.2)$$

где:

– n – количество наблюдений в тестовой выборке,

– $L(f(x_i), y_i)$ – значение функции потерь для i-го наблюдения;

Наиболее распространёнными функциями потерь являются среднеквадратичная ошибка (MSE), средняя абсолютная ошибка (MAE) или доля неправильных классификаций (в задачах классификации).

Для более точной и устойчивой оценки качества модели используется кросс-валидация – метод, при котором выборка многократно делится на обучающую и тестовую части, а результаты усредняются. Это позволяет снизить зависимость итоговой оценки от случайного разбиения данных и получить более объективную характеристику обобщающей способности.

Таким образом, метод оценивания обобщающей способности является ключевым элементом процесса машинного обучения и статистического моделирования. Он обеспечивает объективную проверку надёжности и устойчивости моделей, служит основой для выбора оптимальных гиперпараметров и сравнения различных алгоритмов, а также играет важную роль в повышении достоверности аналитических выводов.

2. АЛГОРИТМ РЕАЛИЗАЦИИ МЕТОДА

Реализация метода оценивания обобщающей способности основывается на последовательности шагов, направленных на объективную проверку качества модели и выявление степени её устойчивости к новым данным. Алгоритм строится вокруг идеи разбиения исходных данных на обучающие и тестовые подмножества, а также анализа ошибки модели на каждом этапе обучения и валидации.

Ниже приведён общий алгоритм оценки обобщающей способности модели:

1. Подготовка данных. Исходный набор данных очищается от пропусков, выбросов и несоответствий. При необходимости выполняется нормализация признаков, кодирование категориальных переменных и случайное перемешивание строк для исключения систематических искажений,

2. Разделение выборки. Данные разбиваются на две или три подвыборки:

   a. обучающую (training set) — используется для построения модели,

   b. валидационную (validation set) — применяется для подбора гиперпараметров (необязательно, но желательно при сложных моделях),

   c. тестовую (test set) — служит для окончательной проверки обобщающей, способности. Наиболее часто применяется пропорция 70 % / 30 % или 80 % / 20 % от общего объёма данных;

3. Обучение модели. На обучающей выборке проводится подбор параметров модели β или весов w (в зависимости от типа алгоритма). В этом процессе минимизируется эмпирический риск, то есть ошибка на обучающих данных [2] (2.1):

$$\hat{E}_{train} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} L(f(x_i), y_i). \quad (2.1)$$

где $L(f(x_i), y_i)$ — функция потерь, выбранная в зависимости от задачи (например, среднеквадратичная ошибка, MAE, accuracy и т.д.).

4. Проверка на тестовых данных. После завершения обучения модель тестируется на ранее неиспользованных данных. Рассчитывается тестовая ошибка (2.2):

$$\hat{E}_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} L(f(x_i), y_i), \quad (2.2)$$

где $L(f(x_i), y_i)$ — функция потерь, выбранная в зависимости от задачи (например, среднеквадратичная ошибка, MAE, accuracy и т.д.).

5. Сравнение обучающей и тестовой ошибок. Если E train значительно меньше E test модель переобучена. Если обе ошибки велики, модель недообучена. Цель — достичь минимальной тестовой ошибки при умеренной сложности модели.

6. Для повышения точности оценки и уменьшения зависимости результата от случайного разбиения данных используется метод k-fold кросс-валидации. Алгоритм кросс-валидации заключается в следующем:

– исходные данные делятся на k равных частей (блоков),

– на каждом шаге k−1 блок используется для обучения, а оставшийся – для тестирования,

– итоговая ошибка вычисляется как среднее значение ошибок по всем k итерациям по формуле [3] (2.3):

$$\hat{E}_{cv} = \frac{1}{k} \sum_{i=1}^{k} \hat{E}_i. \quad (2.3)$$

Типичное значение k варьируется от 5 до 10.

7. Интерпретация результатов. По итогам оценки определяется, насколько модель способна к обобщению. При необходимости выполняется дополнительная настройка гиперпараметров, выбор другой модели или применение методов регуляризации.

Таким образом, алгоритм реализации метода оценивания обобщающей способности является универсальным инструментом анализа качества модели. Он позволяет выявлять переобучение, оптимизировать структуру модели и обеспечивать достоверную оценку её эффективности при работе с новыми данными.

3. КРОССВАЛИДАЦИЯ

Кросс-валидация представляет собой один из наиболее эффективных и широко применяемых методов оценки обобщающей способности модели. Основная идея данного метода заключается в многократном разделении исходных данных на обучающие и тестовые подмножества с последующим усреднением результатов. Это позволяет получить более надёжную и устойчивую оценку качества модели, уменьшая влияние случайного выбора обучающей и тестовой выборок.

Метод кросс-валидации особенно актуален в случаях, когда объём доступных данных ограничен, и выделение отдельной тестовой выборки может привести к потере ценной информации для обучения. Повторяя процесс обучения и проверки на различных подмножествах, кросс-валидация обеспечивает максимально эффективное использование данных.

Рассмотрим основные принципы кросс-валидации. Пусть имеется набор данных, представленный ниже (3.1):

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{3.1}$$

где

– x_i – вектор признаков,

– y_i – значение целевой переменной.

Метод кросс-валидации заключается в следующем: набор данных D случайным образом делится на k непересекающихся подмножеств (блоков) одинакового размера. На каждом шаге i из выборки исключается один блок D_i, который используется в качестве тестовой выборки, а оставшиеся k-1 блоков объединяются в обучающую выборку. Модель обучается на обучающих данных и тестируется на D_i, рассчитывается ошибка E_i. После выполнения всех k итераций вычисляется среднее значение ошибки по всем блокам (3.2):

$$\hat{E}_{cv} = \frac{1}{k} \sum_{i=1}^{k} \hat{E}_i. \tag{3.2}$$

Это значение и используется как итоговая оценка обобщающей способности модели.

Существует несколько разновидностей метода кросс-валидации, отличающихся способом деления данных и числом итераций.

Например, простая (k-fold) кросс-валидация – это наиболее распространённый вариант, при котором выборка делится на k равных частей.

Типичные значения k – от 5 до 10. Метод обеспечивает хорошее соотношение между точностью оценки и вычислительными затратами.

Leave-One-Out (LOO) кросс-валидация – частный случай k-fold при k=N, то есть каждый объект поочерёдно используется как тестовый, а оставшиеся N-1 для обучения. Формула для итоговой ошибки (3.3):

$$\hat{E}_{loo} = \frac{1}{N} \sum_{i=1}^{N} L(f_{-i}(x_i), y_i) \quad (3.3)$$

Stratified k-fold (стратифицированная кросс-валидация) используется в задачах классификации, когда важно, чтобы доли классов в каждом подмножестве сохраняли исходное соотношение. Это предотвращает смещение оценки при несбалансированных данных.

Repeated k-fold (повторная кросс-валидация) представляет собой усреднение результатов нескольких независимых k-fold процедур с разным случайным разбиением данных. Позволяет дополнительно снизить дисперсию оценки качества.

Time Series Cross-Validation (валидация временных рядов) применяется для временных данных, где порядок наблюдений имеет значение. Данные не перемешиваются, а разделяются с сохранением временной последовательности, чтобы не допустить «утечки будущего» (information leakage).

4. СРАВНЕНИЯ С ДРУГИМИ МЕТОДАМИ АНАЛИЗА

Методы оценивания обобщающей способности и, в частности, кросс-валидация занимают важное место среди инструментов анализа качества моделей машинного обучения. Однако существуют и другие подходы, позволяющие проверять надёжность и точность моделей. Сравнение этих методов позволяет лучше понять преимущества и ограничения кросс-валидации, а также выбрать оптимальный способ оценки для конкретной задачи.

Сравнение с методом простого разделения выборки (Hold-Out). Наиболее простым способом проверки обобщающей способности является метод hold-out, при котором выборка делится на две части: обучающую и тестовую. Оценка качества модели проводится по ошибке на тестовой выборке (4.1):

$$\hat{E}_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} L(f(x_i), y_i). \quad (4.1)$$

Метод прост в реализации и требует минимальных вычислительных затрат, однако его результат сильно зависит от конкретного разбиения данных. При малых объёмах выборки полученная оценка может быть нестабильной и не отражать истинную способность модели к обобщению.

В отличие от hold-out, кросс-валидация проводит многократное разбиение данных, обеспечивая усреднение ошибок по нескольким итерациям. Это значительно повышает точность оценки и снижает влияние случайных факторов, что делает её предпочтительным выбором при ограниченных данных.

Сравнение с бутстреп-методом (Bootstrap). Другим популярным методом оценки качества модели является бутстреп (bootstrap), основанный на случайном повторном выборочном извлечении данных с возвращением. Из исходного набора формируется множество псевдовыборок, на каждой из которых модель обучается и проверяется. Средняя ошибка по всем выборкам используется как оценка обобщающей способности (4.2):

$$\hat{E}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{E}_b, \quad (4.2)$$

где

– B – количество бутстреп-выборок.

Главное преимущество бутстрепа заключается в возможности оценивать не только среднее качество модели, но и дисперсию (стабильность) оценок. Однако данный метод может давать смещённые результаты при наличии зависимостей между наблюдениями или при малом размере исходных данных.

Бутстреп и кросс-валидация представляют собой два метода оценивания обобщающей способности моделей, однако они различаются по своему принципу и практическим свойствам. Метод бутстрепа основан на многократной случайной выборке объектов из исходных данных с возвращением, что позволяет получать различные обучающие подвыборки и оценивать не только точность, но и дисперсию оценки. При этом часть объектов может не попасть в выборку, а другие — повторяться, что вносит определённое смещение в результаты. Напротив, метод кросс-валидации использует разделение данных на k непересекающихся частей без возвращения. Каждая часть последовательно служит тестовой, а остальные — обучающими, что обеспечивает минимальное смещение оценки и эффективное использование всего набора данных. С вычислительной точки зрения бутстреп является более затратным, поскольку требует большого числа повторных обучений модели, тогда как кросс-валидация имеет среднюю вычислительную сложность и чаще применяется на практике для оценки обобщающей способности алгоритмов машинного обучения.

Сравнение с информационными критериями (AIC, BIC).

В задачах статистического моделирования часто применяются информационные критерии, такие как AIC (Akaike Information Criterion) и BIC (Bayesian Information Criterion). Информационные критерии удобны для сравнения нескольких моделей, но они предполагают, что данные соответствуют выбранной вероятностной модели. В отличие от них, кросс-валидация не требует предположений о распределении данных и может применяться к любым типам моделей, включая нелинейные и не параметрические. Эти методы не требуют разбиения данных, а оценивают модель по соотношению точности подгонки и количества параметров (4.3):

$$AIC = 2k - 2\ln(L), \quad BIC = k\ln(n) - 2\ln(L), \tag{4.3}$$

где

– k – число параметров модели,

– L – значение функции правдоподобия,

– n – объём выборки.

Сравнение показывает, что кросс-валидация является универсальным и надёжным методом оценки обобщающей способности моделей. В отличие от простого разделения выборки, она даёт устойчивые результаты; по сравнению с бутстрепом — менее ресурсоёмка; а в отличие от информационных критериев — не требует строгих статистических предположений.

Таким образом, кросс-валидация занимает промежуточное место между статистическими и эмпирическими подходами, обеспечивая оптимальный баланс между точностью, вычислительной сложностью и универсальностью применения.

5. СИЛЬНЫЕ И СЛАБЫЕ СТОРОНЫ МЕТОДА

Метод кросс-валидации обладает рядом существенных преимуществ, благодаря которым он широко применяется в машинном обучении и статистическом анализе. Он обеспечивает более надёжную и стабильную оценку обобщающей способности модели, так как использует разные подмножества данных для обучения и тестирования. Кроме того, кросс-валидация позволяет эффективно использовать весь доступный набор данных, что особенно важно при ограниченном объёме выборки, и предоставляет возможность объективного сравнения различных моделей и их гиперпараметров.

Однако метод имеет и определённые ограничения. Его основным недостатком является высокая вычислительная сложность, особенно при работе с большими наборами данных или сложными моделями. Также при неверной стратификации может возникнуть смещение оценки, а в задачах, связанных с временными рядами, применение стандартных схем разбиения данных может привести к нарушению временной последовательности, что требует использования специализированных подходов.

В целом, кросс-валидация представляет собой универсальный и объективный инструмент оценки качества моделей, позволяющий выявлять переобучение и принимать обоснованные решения при выборе оптимальной модели для конкретной задачи.

6. АНАЛИЗ РЕЗУЛЬТАТОВ И ОБСУЖДЕНИЕ

Анализ результатов оценки обобщающей способности модели с использованием кросс-валидации позволяет сделать выводы о её качестве и устойчивости. При проведении экспериментов каждая итерация кросс-валидации даёт значение ошибки на тестовом подмножестве, а усреднение этих значений позволяет получить объективную оценку производительности модели. Если разброс ошибок между итерациями невелик, это указывает на стабильность и хорошую способность модели к обобщению. Напротив, значительные колебания ошибок свидетельствуют о чувствительности модели к составу обучающих данных, что может говорить о переобучении или недостаточной сложности модели.

Сравнение результатов разных моделей или вариантов гиперпараметров по средней ошибке кросс-валидации позволяет определить наиболее эффективную конфигурацию. Например, при снижении ошибки на обучающей выборке, но росте ошибки на тестовой можно заключить, что модель переобучена. Если же обе ошибки остаются большими, следует рассмотреть возможность увеличения сложности модели, изменения признаков или выбора другого алгоритма.

Таким образом, результаты кросс-валидации служат не только показателем качества, но и инструментом диагностики модели, позволяя выявлять проблемы с переобучением, подбирать оптимальные параметры и обеспечивать высокую точность прогнозирования на новых данных.

ЗАКЛЮЧЕНИЕ

В ходе исследования были рассмотрены основные методы оценивания обобщающей способности моделей, среди которых особое внимание уделено методу кросс-валидации. Этот подход позволяет получить объективную и устойчивую оценку качества модели, обеспечивая сбалансированное использование доступных данных и снижение влияния случайных факторов.

Кросс-валидация доказала свою эффективность при решении задач машинного обучения, статистического анализа и моделирования, особенно в условиях ограниченных выборок. Благодаря многократному разделению данных и усреднению результатов, метод позволяет выявлять переобучение, сравнивать различные модели и выбирать оптимальные параметры, что делает его универсальным инструментом оценки.

Тем не менее, применение кросс-валидации требует значительных вычислительных ресурсов и аккуратного подхода к стратификации данных, особенно при анализе временных рядов или несбалансированных выборок. Несмотря на эти ограничения, метод остаётся одним из наиболее надёжных и широко используемых средств проверки обобщающей способности моделей.

Таким образом, кросс-валидация играет ключевую роль в современном анализе данных и машинном обучении, обеспечивая достоверную оценку производительности моделей и повышение качества принимаемых решений.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie [Электронный ресурс] – Режим доступа: https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf (дата обращения: 05.10.2025).

2. An Introduction to Statistical Learning with Applications in R Gareth James [Электронный ресурс]. – Режим доступа: https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf (дата обращения: 05.10.2025).

3. Pattern Recognition and Machine Learning Christopher M. Bishop [Электронный ресурс]. – Режим доступа: https://github.com/Benlau93/Data-Science-Curriculum/blob/master/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf (дата обращения: 05.10.2025).